(12)

PROCEDURAL DEBIASING

Lola L. Lopes

WHIPP #15 October 1982

This research was supported by the Engineering Psychology Programs, Office of Naval Research, ONR Contract N00014-81-C-0069 Work Unit NR197-C50.

WISCONSIN HUMAN INFORMATION PROCESSING PROGRAM

DEPARTMENT OF PSYCHOLOGY • UNIVERSITY OF WISCONSIN • MADISON, WISCONSIN 53706

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER WHIPP 15	2. GOVT ACCESSION NO. AD-A120844	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCEDURAL DEBIASING		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report 1 Jan 82 -- 1 Oct 82 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Lola L. Lopes		8. CONTRACT OR GRANT NUMBER(s) N00014-81-C-0069
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Psychology University of Wisconsin Madison, WI 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR197-068
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research 800 North Quincy Street Arlington, Virginia 22217		12. REPORT DATE October 1982 13. NUMBER OF PAGES 39
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; Distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayesian inference
debiasing
averaging
conservatism

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

As knowledge increases about human judgment processes, it is natural to suppose that it will be possible to use this knowledge in order to improve human judgment in situations where biases of various sorts have been shown to occur. Despite the reasonableness of this expectation, judgmental debiasing has proven extraordinarily difficult in most cases. This paper suggests that the reason for this failure is that debiasing must be in terms of the procedures that are actually used in the act of judging, procedures about which very little is known. Two experiments are presented that

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

illustrate how such procedural debiasing can be used to debias a Bayesian inference task. In the first experiment, a training procedure is used that corrects a common error in the direction of the adjustment process that subjects use when integrating later evidence with earlier partial judgments. In the second procedure a focusing technique is used to improve the relative weighting of samples in the overall judgment. Each of the procedures accomplishes its particular end, and taken together the two procedures allow naive subjects to produce judgments that are essentially Bayesian. These results are discussed in terms of a theoretical model of the judgment process in which four basic stages are repeated cyclically: (a) initial scanning of the stimulus information; (b) selection of items for processing in order of importance; (c) extraction of scale values on the given dimension of judgment; and (d) adjustment of a composite value that summarizes already-processed components.

Procedural Debiasing

Lola L. Lopes
University of Wisconsin--Madison

As increasingly more is known about human judgment processes, it becomes reasonable to expect that this knowledge can be used to help people make better judgments. This is particularly true in situations where failures of judgment seem to be orderly manifestations of the processing mechanisms used by the judge and not merely the random errors that might be attributed to inattention, insufficient knowledge, faulty memory and similar nonsystematic factors. Unfortunately, however, it has been easier to imagine such improvement than to produce it (Fischhoff, 1982).

Probably the first attempts at debiasing human judgments were aimed at reducing the tendency of naive subjects in Bayesian inference tasks to produce judgments that are "conservative" relative to the Bayesian norm (Edwards, 1968). In discussing these early debiasing efforts, it is useful to rely on a classification scheme devised by Fischhoff (1982) in which debiasing procedures are categorized according to whether they lay the blame for the bias at "the doorstep of the judge, the task, or some mismatch between the two" (p. 424).

Allegations that a task is faulty generally center on the possible failure of the experimenter to instill in subjects sufficient understanding of the task and sufficient motivation for proper performance. In the case of Bayesian inference, Phillips and Edwards (1966) used specialized payoff schemes and feedback in order both to encourage subjects to try harder and to help them better understand the task. Generally speaking, these methods had some effect in reducing conservatism, but they were not able to eliminate it.

A second task fault that was investigated involved a potential bias in the response scale. The argument ran that "correct" performance in Bayesian tasks often requires the production of extreme responses, particularly when the judgments must be given on a probability scale. If subjects are hesitant to make such extreme judgments, conservatism can result. Phillips and Edwards (1966) tested this hypothesis by comparing judgments on probability scales with judgments on "odds" and "log odds" scales which require less extreme responding. They found that use of response scales such as these reduced conservatism only slightly relative to the more conventional probability scale.

A more recent attempt at debiasing falls in Fischhoff's (1982) category of attributing the bias to a mismatch between the task and the judge. Eils, Seaver, and Edwards (1977) based their procedure on the observation that the judgments of naive subjects in Bayesian tasks are often more like averages (or estimates of population proportion) than like inferences, (Beach, Wise, & Barclay, 1970; Marks & Clarkson, 1972, 1973; Shanteau, 1970, 1972). This being so, Eils et al. hypothesized that subjects might be better at judging the mean log likelihood ratio for a set of samples than at judging the more standard cumulative log likelihood ratio. They also noted that the averaging response would reduce problems of "response bias" if there were any operating.

The hypothesis was tested by using two groups of subjects, one of which rated their average certainty for the target hypotheses and the other of which rated their cumulative certainty. Responses from both groups were then converted to log posterior odds form and regression analysis was performed for each subject comparing inferred log posterior odds to veridical log posterior odds. The results supported the hypothesis: log odds inferred from average certainty judgments were definitely closer to veridical than odds inferred from cumulative certainty judgments.

The present research represents an attempt at debiasing that falls in Fischhoff's remaining category, that of attributing error to faulty judges. Like the work of Eils et al. (1977), the research begins with the observation that untutored subjects in Bayesian tasks tend to produce data that are more like averages than like inferences. But unlike the approach of Eils et al., no attempt is made to "engineer" the task to be better suited to human proclivities. Instead, debiasing involves (a) analyzing the procedures that untutored subjects use when they produce averages, (b) warning subjects about the specific procedures that are inappropriate, and (c) providing subjects with appropriate procedures that can be used in place of the inappropriate procedures.

Averaging and Adjustment in Bayesian Inference

Bayesian inference tasks are usually instantiated in terms of the "bookbags and poker chips" paradigm in which subjects consider two well-specified hypotheses (i.e., bookbags) usually involving populations of binary events (i.e., red and blue poker chips). Typically the subject is shown two or more samples, often sequentially, and is asked after each sample to rate the strength of his

or her belief about which population generated the samples.

According to Bayes' theorem, the normative response for such situations is found by multiplying the prior odds ratio for the two hypotheses by the likelihood ratio of the sample data given the two hypotheses. This yields the posterior odds ratio:

$$\frac{p(H1|D)}{p(H2|D)} = \frac{p(D|H1)}{p(D|H2)} \times \frac{p(H1)}{p(H2)} \qquad (1)$$

Alternatively, one can write Bayes' theorem to give the probability of a particular hypothesis:

$$p(H1|D) = \frac{p(D|H1) \cdot p(H1)}{p(D|H1) \cdot p(H1) + p(D|H2) \cdot p(H2)} \qquad (2)$$

Note that in these equations, the relationship between current data and previous data is multiplicative.

How do naive humans perform when they are asked to provide inferences in Bayesian tasks? As has already been mentioned, human inferences differ from Bayesian inferences in two important ways: (a) the individual judgments are typically conservative relative to the Bayesian norm, and (b) the pattern of judgments is suggestive more of averaging or estimation than of inference (Beach, Wise, & Barclay, 1970; Marks & Clarkson, 1972, 1973; Shanteau, 1970, 1972). Shanteau (1970) hypothesized that people's judgments in such tasks could be modeled by an algebraic rule in which the response, $\underline{R}$, at any serial position, $\underline{n}$, is given by a weighted average of the scale values, $\underline{s_i}$, of the previous and current sample events:

$$\underline{R_n} = \sum_{\underline{i}=0}^{\underline{n}} \underline{w_i s_i} \qquad (3)$$

In this equation the $\underline{w_i}$ are weights that sum to unity and the term $\underline{w_0 s_0}$ signifies the weight and scale value of a neutral initial impression. It should be noted that averaging is necessarily conservative relative to inference because averages always lie within the range of the component stimulus values whereas inferences are often more extreme than any of their component values.

Shanteau's model is successful in accounting for the quantitative features of the data, but it does not suggest either why or how averaging occurs. In previous research (Lopes, 1981; Lopes & Johnson, 1982; Lopes & Oden, 1980) I have suggested that averaging may occur because subjects integrate the stimulus information serially via an "anchoring and adjustment" process (Tversky & Kahneman, 1974). In this process subjects are hypothesized to integrate "new" information into "old" composite judgments by adjusting the old value as necessary to make the new composite lie somewhere between the old composite and the value of the new information. Althcugh this process is qualitatively equivalent to averaging, it does not presuppose that subjects ever "compute" an average in any algebraic -- or even any conscious -- sense of the term. Instead, averaging is simply the natural consequence of the adjustment procedure.

One prediction of the adjustment model is that subjects in the Bayesian task will occasionally make adjustments that are strictly in the wrong direction. Consider two samples, both of which support the same hypothesis but to different degrees. If a subject is first shown the weaker sample, we suppose that some weak preliminary judgment will be made in favor of the supported hypothesis. When the subject is later shown the stronger sample, adjustment will be made in the direction of increased support for the hypothesis. This is entirely appropriate qualitatively. But if the samples are reversed so that the weaker sample follows the stronger, qualitatively inappropriate adjustment ought to result. That is, the preliminary judgment ought to produce a relatively strong result. When the weaker sample is later integrated into the judgment, adjustment should be in the neutral direction since the value of the weaker sample is more neutral than the preliminary judgment. Such adjustment is obviously inappropriate since movement in the neutral direction is de facto movement toward the alternative or non-supported hypothesis.

Previous research (Lopes, 1981) has clearly supported the prediction that subjects will adjust in the normatively incorrect direction when a weaker sample favoring some particular hypothesis follows a stronger sample favoring the same hypothesis. The present research is aimed at finding out whether these "directional errors" can be eliminated by training that warns subjects of the occurrence of the errors and also teaches subjects an alternative procedure that is directionally correct.

Two experiments are presented. The first experiment focuses on improving subjects' adjustment procedures qualitatively in specific cases where adjustment

errors are known to occur. The second experiment extends the training to include instruction in a selectional procedure that is hypothesized to improve subjects' quantitative performance.

Experiment 1

Method

Experimental task. Subjects in both conditions were asked to put themselves in the place of a machinist whose job is to make decisions concerning the maintenance of milling machines using samples of parts produced by the machines. The judgment concerns whether or not a critical spring has broken inside the machine. Subjects were told that normal machines have a rejection rate of about 12 parts per 1000 parts produced (H12/1000), whereas machines with broken springs have a rejection rate of about 20 parts per 1000 (H20/1000). Thus, in abstract terms, the subjects were required to decide between alternate Bernoulli processes, one with $p = .012$ and the other with $p = .02$, with p being the probability of a rejected part.

Stimulus design. The stimulus design was a 9 x 9, first-sample x second-sample, factorial design in which the levels of both factors comprised the same samples of parts. These were 12, 13, 14, 15, 16, 17, 18, 19, and 20 rejects per 1000 parts, respectively.

Procedure. Subjects were run individually in sessions that took about 40 minutes for control subjects and 50 minutes for trained subjects. At the beginning of the session subjects were brought into a sound proof booth and seated in front of a computer controlled video terminal. Control subjects were then given general instructions about the nature of the task and shown how to read the stimulus display. A sample of a stimulus display is shown in Figure 1. At the top of the display is a box showing a sample with 13 rejects out of 1000 parts. Under this is a notation showing that this is the first of two samples. At the bottom of the display is a response scale anchored at the left by the words "machine normal" and at the right by the words "machine faulty".

- - - - - - - - - - - - - - -
Figure 1 about here
- - - - - - - - - - - - - - -

The procedure for each trial was identical. Subjects read the information for the first sample and then rated their degree of belief as to whether the

machine was milling normally or not. They made their ratings using a joystick to move the rating arrow (shown in the middle of the scale in Figure 1) along the response scale. When they finished their initial rating, subjects pushed a button on the response box. This caused the initial rating to be transmitted to the computer and also caused the first sample to be replaced by a second sample of parts from the same machine. Subjects revised their initial rating to account for the new sample and pushed the response button to transmit their final response to the computer. Then they initiated the next trial by returning the response arrow to the middle of the scale and pushing the button again.

The instructions for trained subjects were essentially identical to those for control subjects through the explanation of the stimulus display and the rating response, except that trained subjects were told at the outset that they would be taught a procedure for avoiding a common judgment error. The actual training took place during the early practice trials. The first practice trial was a weak-strong pair (17/19) that was chosen especially to elicit correct responses from all subjects. For this trial, all subjects initially rated a sample of 17 rejects to favor the faulty machine moderately and then adjusted this rating to favor the faulty machine even more strongly after presentation of the sample of 19 rejects.

The second trial was a strong-weak pair (13/14) chosen to elicit the directional error. On this trial subjects were shown the first sample (13 rejects) and allowed to make their initial rating and to transmit their response. Then they were shown the second sample (14 rejects) and were allowed to make their adjustment, but they were stopped before they transmitted the response. Most of the subjects (20 of 31) made their adjustment in the wrong direction and were read the instructions reproduced below. The others were read similar instructions, but with wording changed to accommodate the fact that they had, in fact, responded correctly on this trial.

> Before you transmit your response, let me talk with you about your response. You shouldn't feel bad, but remember I told you that many people make an error in this task. Well, you just made it. Let me explain it to you. Most people, if they are given a sample of <u>14</u> rejected parts as a first sample, say that the machine is more likely to be functioning normally than not. But when they are given a sample of 14 rejected parts <u>after</u> they have just been given a sample of <u>13</u> rejected parts,

> they tend to adjust their judgment toward the right, that is, toward the machine being broken. Now if you think about it, this is an error of adjustment since a sample of 14 rejects favors the normal machine and therefore provides additional evidence that the machine is normal. Thus, the adjustment should be toward the left, that is, toward the machine functioning normally. Do you understand this so far?

After subjects indicated that they understood what the error was, the experimenter taught them a simple procedure for avoiding the error. Basically, this was to separate each judgment operation into two steps: (a) the labeling of each sample as either favoring the "normal" hypothesis or the "faulty" hypothesis and (b) the adjustment of the current response in the direction given by the label. (It is convenient to think of the initial rating produced by the subject after presentation of the first sample as involving an adjustment made to an earlier and implicit "neutral" response produced by the subject at the onset of each new trial.) Thus, when both first and second samples favored the same hypothesis, both the initial rating and the final adjustment would be made in the same direction relative to the neutral point, and only when the second sample favored a hypothesis different from the first would the final adjustment be opposite in direction to the initial rating.

After teaching subjects the judgment procedure, the experimenter asked them to respond to several trials on their own, while verbalizing what they were doing. This allowed the experimenter to check that they were explicitly separating the labeling and the adjustment steps and that they were adjusting at each step in the direction given by the labeling operation. Among these training trials were two for which these samples were identical. When the first such trial (17/17) appeared, the experimenter waited to see whether the subject would adjust for the second sample and then stopped the trial for further instruction. Subjects who had failed to adjust (8 out of 31) were told, "Now this kind of trial also causes errors. Let me explain. Your first sample was 17 rejects and you judged the machine as likely to be broken. Then you got new evidence of 17 rejects also favoring the machine being broken, but you didn't adjust. Actually you should have adjusted since that is additional evidence in favor of the machine being broken. Do you see what I mean?" Subjects who had adjusted correctly were read similar instructions, but modified to accord with their correct response.

It is important to note that the training procedure involved only qualitative features of the judgment process. At no time were subjects given instruction concerning how they ought to evaluate the sample information quantitatively. Although such training might be helpful generally, the aim of the present research was to determine the degree to which judgment can be improved by strictly procedural means, that is, by giving subjects better procedures for operating on information rather than by giving them better or more accurate information.

Altogether there were 13 trials for practice and training for the trained subjects. Control subjects received the same 13 trials for practice, but with no training. Then both groups of subjects received two replications of the stimulus design, bringing the total number of trials to 175 per subject. Experimental trials within each replication were ordered randomly but with the restriction that no sample appear either as first-sample or second-sample on two consecutive trials.

Subjects. The subjects for control and trained groups were, respectively, 30 and 31 student volunteers from the University of Wisconsin-Madison. Approximately half were males and half females. They served for credit to be applied to their course grades in introductory psychology.

Results and Discussion

Two questions are of interest in this experiment. The first is whether training concerning directional adjustment errors can prevent or at least reduce their prevalence in the inference task. The second is whether, given that such prevention or reduction of errors is possible, this leads to improvement in the accuracy of the final judgments.

Data bearing on the first question are given in Table 1. Five subjects have been dropped from the control condition and five from the trained condition since these subjects appeared to base their final judgments entirely on the second sample. Note, however, that the basic results of the experiment would have been the same whether these subjects were retained or not.

Subjects were unanimous in treating samples of 12 to 15 rejects per 1000 parts as favoring the machine being normal and samples of 17 to 20 rejects as favoring the machine being broken, but they were highly variable in how they treated samples of 16 rejects. (Actually, such samples favor slightly the machine being broken.) Some subjects tended to treat these as neutral, others

treated them as favoring one or the other hypothesis, and still others treated then inconsistently across trials. For reasons of this variability, pairs involving 16 rejects are not considered explicitly in the formal analysis. However, an interesting problem involving these samples that occurred for some subjects is described in the General Discussion.

- - - - - - - - - - - - - - -

Table 1 about here

- - - - - - - - - - - - - - -

Taken together, there were 20 pairs in which adjustment errors might have been expected. These comprised the eight pairs along the diagonal of the stimulus design in which the two samples are identical (i.e., 12/12, 13/13, 14/14, 15/15, 17/17, 18/18, 19/19, 20/20) and the twelve non-diagonal pairs in which a (stronger) sample favoring a particular hypothesis is followed by a weaker sample favoring the same hypothesis (i.e., 12/13, 12/14, 12/15, 13/14, 13/15, 14/15, 20/19, 20/18, 20/17, 19/18, 19/17, 18/17). These pairs are indicated in the table as "diagonal" pairs and "strong-weak" pairs, respectively. The table also gives results for the set of "weak-strong" pairs. These are exactly the same set as the strong-weak pairs except that the stronger sample in each pair is preceded by the weaker. Since for these pairs the intuitive direction of adjustment is normatively correct, they provide an estimate of the rate of adjustment errors that occur for reasons other than the incompatibility of the normative response with the intuitive direction of adjustment (i.e., misreading the stimulus).

Looking first at strong-weak pairs and weak-strong pairs, it is clear that the training procedure has been effective in reducing the number of directional adjustment errors, where "error" refers to an explicit adjustment in the nonnormative direction. (Including as errors occasions on which no adjustment was made would have produced essentially the same results.) For the control group there is an average of 13.4 errors per subject (out of 24 maximum) for strong-weak pairs compared to an average of only .40 errors per subject for weak-strong pairs; $F(1,24) = 60.43$, $p < .05$). For the trained subjects, however, the average is 3.11 errors for strong-weak pairs compared to .42 errors for weak-strong pairs; $F(1,25) = 8.33$, $p < .05$. Comparing across groups, the trained subjects have significantly fewer errors then control subjects for strong-weak pairs [$F(1,49) = 113.46$, $p < .05$] but not for weak-strong pairs [$F < 1$].

The final row of the table gives the results for diagonal pairs. Adjustment errors have been scored for these pairs only if there was room for the adjustment to occur (i.e., the response was not already at the end of the scale) and if there

was either no adjustment at all or adjustment in the wrong direction. (Errors of the latter type were very rare.) Mean errors (out of a maximum of 16) were 3.24 for the control group and 1.08 for the trained group. Both these error rates are significantly different from zero [$F(1,24) = 26.18$ and $F(1,25) = 8.99$, respectively, $p < .05$], and the rate for the control group is significantly higher than that for the trained group [$F(1,49) = 10.29$, $p < .05$].

In general, it appears that the training procedure was able to reduce (although not completely to eliminate) directional adjustment errors, particularly for strong-weak pairs. The question remains, however, as to whether this reduction was accompanied by improved accuracy of judgment (i.e., reduced conservatism). Figure 2 gives the final judgment data for the control group pooled over both subjects and replications. For purposes of comparison, Figure 3 gives theoretical values for an optimal Bayesian judge. In Figure 2, the data for pairs where errors are likely (i.e., strong-weak pairs and diagonal pairs) are shown by filled symbols and the data for remaining pairs are shown by open symbols. The row parameter in both cases is number of rejects in the second sample.

- - - - - - - - - - - - - - - -
Figures 2 and 3 about here
- - - - - - - - - - - - - - - -

It is clear graphically that there is a large difference between the data pattern produced by the control subjects and the theoretical pattern: The theoretical data have a "barrel" shape whereas the control data look more like a set of parallel lines. This appearance is borne out by analysis of variance: Although the data for control subjects have a significant interaction [$F(64,1536) = 2.66$, $p < .05$), it accounts for only .7% of the total systematic sum of squares. By way of contrast, analysis of variance on the theoretical values indicates that the interaction should account for 4.66% of the systematic sum of squares.

The data for the trained subjects are in Figure 4. Overall, the figure presents the same appearance as that for the control group, although the interaction term [$F(64,1600) = 4.98$, $p < .05$) is somewhat larger, accounting for 1.2% of the systematic sum of squares.

- - - - - - - - - - - - - - - -
Figure 4 about here
- - - - - - - - - - - - - - - -

Although Figure 4 gives all the data, the points that are critical for the training procedure are just those that are filled. Comparison of these

critical pairs for control and trained subjects shows that subjects who had received training were, indeed, more accurate in their judgments for these points. Figured on group means, the root-mean-squared-deviations between obtained and theoretical were, for the control group, .0723 for strong-weak pairs and .0227 for diagonal pairs, relative to the 0-1 response scale. For the trained subjects, however, these values were .0187 and .0047, respectively.

The data for the critical pairs are about what might be expected given the nature of the training, but an unexpected finding is that improved performance on strong-weak pairs generalized to the corresponding weak-strong pairs: Although the training procedure did not in any way attempt to modify subjects' procedures for judging weak-strong pairs, trained subjects did about as well on these (RMSD = .0220) as they did on the strong-weak pairs. In the same way, the control subjects did about as poorly on weak-strong pairs, RMSD = .0626, as they did on the strong-weak pairs.

This generalization of improved accuracy from strong-weak to weak-strong pairs is of interest since it suggests that the training instructions may have been effective not only in helping subjects avoid the specific adjustment error, but also in helping them understand the task better. Although the present data do not speak to the issue directly, previous data showing that the judgments of naive subjects are more like estimates of population proportion than they are like inferences (Beach, Wise, & Barclay, 1970; Marks & Clarkson, 1972, 1973; Shanteau, 1970, 1972) suggests that subjects may have difficulty understanding the difference between inference and estimation. By focusing attention on the directional errors in inference that occur for strong-weak pairs, one may also, by serendipity, focus attention on the special characteristics that distinguish inference from estimation and hence, improve subjects' understanding of the task.

But if trained subjects do understand the inference process better than control subjects, why do their data show the same tendency toward parallelism? Put another way, why are their inferences so conservative for those heterogeneous pairs (shown in the upper left and lower right quadrants of the figures) in which the two samples favor different hypotheses? The answer to this may lie in the weights that subjects give to the various samples.

Consider a situation that is like the current one except that subjects are actually <u>instructed</u> to estimate the proportion of rejected parts for a particular machine. If the two samples are of equal size and equal reliability,

the subject ought to give them equal weight and simply average the values. Furthermore, no matter what value a particular sample has (i.e., whether the first sample is 12, 14, 16 or any other number of rejects), the value itself should not affect the weight of the sample in the overall judgment. A subject who followed such a "constant weighting" strategy would produce a parallel pattern of data such as is found in Figures 2 and 4.

The Bayesian task, however, requires that subjects adopt a "differential weighting" strategy: Samples that are extreme (i.e., 12 or 20) are more diagnostic than samples that are nearer neutral (i.e., 15 or 17), and should be weighted more heavily in the inference process. But this is what, apparently, subjects do not naturally do in the Bayesian task (i.e., Beach, Wise, & Barclay, 1970; Shanteau, 1970) or in a great many other tasks as well (c.f. Anderson, 1974). Thus, it may be that subjects in the trained condition do understand the Bayesian task better than their control condition analogs, at least in the sense that they are really integrating evidence and not merely integrating sample sizes, but they may not understand that the more extreme estimates are more diagnostic and hence should be accorded greater weight. For homogeneous pairs in which both samples favor the same hypothesis, such a misunderstanding would not be likely to impair accuracy much since subjects' responses are forced to converge (just as they ought to) by the end of the response scale. For heterogeneous pairs, however, the misunderstanding is more serious since there is nothing to prevent subjects from making overly large adjustments given only weakly diagnostic information, thus causing the poor correspondence between theoretical and obtained for these particular pairs.

Experiment 2 investigates whether this hypothesized problem with intuitive weighting of information can be alleviated by a modification of the training procedure used in Experiment 1.

Experiment 2

Method

Task and design. The task for Experiment 2 was exactly like the task for Experiment 1 except that the two samples within each pair were presented simultaneously. The stimulus design was the same as had been used in Experiment 1.

Procedure. The procedure for the control subjects was essentially the same as for Experiment 1 except for the differences occasioned by the simultaneous

stimulus display. However, the instructions for the trained subjects were more detailed and were applied to every kind of stimulus pair. When trained subjects were first brought into the experiment they were told that they would be taught a simple procedure that would allow them to make good judgments in a particular kind of inference task. Then they were told about the task situation (i.e., the machine maintenance problem) and were instructed how to read the stimulus display and use the response device. The actual training began only after it was clear that the subjects understood the stimulus situation.

Subjects were taught a four step procedure to be applied to every stimulus pair. The steps were introduced to subjects and explained as the subjects worked through a series of practice trials. During this training period subjects were asked to work through the steps out loud so that the experimenter could check on their understanding and use of the procedure. The steps were as follows:

(a) Judge for each sample separately whether it supports the "normal" hypothesis or the "faulty" hypothesis or whether it is neutral.

(b) Decide which of the two samples supports its own hypothesis more strongly.

(c) Make an initial rating as to whether the machine is faulty or not based only on the stronger of the two pieces of evidence. If both pieces are equally strong, either can be used as the basis for the initial rating.

(d) Adjust the initial rating in order to take into account the second, (weaker) piece of evidence.

  (i) If the second piece of evidence favors the same hypothesis as the first, then "consider the portion of the response scale between [the] original rating and the [appropriate] end of the scale and move the arrow into this region according to how strong the remaining evidence is."

  (ii) If the second piece of evidence favors the opposite hypothesis, then "consider the portion of the rating scale between [the] original rating and the neutral position and adjust back into this region according to how strongly the sample [supports the other hypothesis]."

Note that although the procedure sounds complicated when summarized, it was much simpler to follow in the context of actual stimulus pairs. No subject appeared to have any great difficulty in following the procedure during training

or in executing the task afterward.

Altogether there were 20 trials for practice and training for the trained subjects. Control subjects received the same 20 practice trials, but with no training. Then both groups of subjects received two replications of the basic stimulus design, bringing the number of trials to 182 per subject. Experimental trials within replication were ordered randomly but with the restriction that no given sample appear on two consecutive trials. The full experiment required about 40 minutes for control subjects and about 55 minutes for trained subjects.

Subjects. The subjects were 56 student volunteers from the University of Wisconsin--Madison, split evenly between the control and the trained conditions. About half were males and half females. Most subjects served for pay, although a few served for credit to be applied to their course grade in introductory psychology.

Results and Discussion

The data for the control subjects are given in Figure 5 pooled over both subjects and replications. Samples designated "first" appeared above the other sample in the simultaneous display.

Note that the pattern of judgments is essentially identical to that of the control subjects in Experiment 1. This visual similarity is confirmed by an analysis of variance showing that the interaction, although significant, $F(64,1728) = 1.81$, $p < .05$, accounts for only .42% of the systematic sum of squares. Calculation of the root-mean-squared-deviations between theoretical and obtained reveals an overall RMSD of .1043 for the entire data array, which breaks down to RMSD's of .0968 for homogeneous cells, .1156 for heterogeneous cells, and .0322 for diagonal cells.

- - - - - - - - - - - - - - - -
Figure 5 about here
- - - - - - - - - - - - - - - -

The data for the trained subjects are in Figure 7. Clearly, the training procedure has been effective in making the subjects' responses more optimal. In terms of analysis of variance, the interaction [$F(64,1728) = 33.89$, $p < .05$] now accounts for 4.55% of the systematic sum of squares compared to the optimal value of 4.66%. The overall RMSD between theoretical and obtained is .0480, which breaks down to RMSD's of .0321 for homogeneous cells, .0567 for heterogeneous cells, and .0077 for diagonal cells. Although deviations for heterogeneous

cells are still somewhat larger than those for homogeneous cells, they are much improved compared to those for the control group. It is also interesting to note that the largest deviations between theoretical and obtained now tend to involve overly radical responses. This is particularly evident for homogeneous pairs in which one sample was 12 rejects and the other was near neutral (15 or 16 rejects). In part these errors reflect the fact that subjects tended to treat the judgment task symmetrically, when samples of 12 rejects actually gave considerably less support to the hypothesis H12/1000 than samples of 20 rejects gave to the hypothesis H20/1000.

- - - - - - - - - - - - - - - -
Figure 6 about here
- - - - - - - - - - - - - - - -

General Discussion

Before proceeding to a discussion of the implications of the present research, it is important to point out exactly what the training procedures did and did not "teach" the subjects. Obviously there would be little interest in showing that subjects can learn to use Bayes' theorem if they are given explicit instruction on how to do so. Debiasing becomes of interest only if it is possible to modify subjects' predilections by procedures that are closer to natural modes of thought than is the rote application of an appropriate normative rule. In other words, the goal is to educate the intuition, not merely to improve the performance.

In Experiment 1, the training procedure taught the subjects only one thing that previously they did not know, namely, that adjustments of the initial rating made after presentation of the second sample should always be in the direction of the hypothesis favored by the second sample. In Experiment 2, the explicit training included the same information about adjustment direction but also taught subjects to process the two samples in order of their apparent relative strength. At no time in either training procedure did the experimenter teach the subjects anything about which samples favored which hypothesis or how strongly they did so, nor did she suggest how diagnostic or "weighty" the samples should be considered to be. These matters of sample evaluation were always left entirely to the subjects.

In light of the limited training to which subjects were exposed, the amount of debiasing that occurred is impressive. In Experiment 1, explicit

training was directed only at the 12 strong-weak pairs and the 8 diagonal pairs. It would have been entirely within reason for subjects' responses to the other 61 pairs to be unaffected by the training since, so far as was indicated, there was nothing wrong with their intuitions concerning such pairs. As it turned out, however, improvement generalized from the strong-weak pairs to analogous weak-strong pairs. Obviously, there is no way to know for sure why this improvement occurred, but a possibility that has appeal is that the training focused subjects' attention on the inferential nature of the task and prevented the apparently common tendency to fall into judging the sample proportion rather than the relative likelihood of the two hypotheses. Thus, trained subjects may have benefitted not only from prior instruction concerning how to prevent a particular error, but also by being forced, so to speak, to better understand what it was they were judging.

There was, however, for some trained subjects an interesting failure of generalization for certain pairs in which a diagnostic sample (i.e., a sample favoring one or the other hypothesis) was followed by a sample that the subject judged to be neutral or nondiagnostic (i.e., a sample of 16 rejects). As was noted earlier, there was considerable variability among subjects in how they evaluated samples of 16 rejects. Nevertheless, 9 control subjects and 16 trained subjects seemed reliably to produce initial ratings of about .50 when a sample of 16 rejects appeared as the first sample. Thus, for these subjects we can assume that such samples were judged to be neutral. When such samples followed diagnostic samples, however, all of the control subjects and all but 5 of the trained subjects adjusted their initial ratings toward neutral, which is normatively inappropriate given that the sample is judged to support neither hypothesis. For the control subjects, such errors are not surprising (c.f. Shanteau, 1975; Troutman & Shanteau, 1977). But the question is why so many trained subjects, if they really <u>understood</u> the task better than control subjects, also made the inappropriate adjustment. The answer may lie in how these subjects interpreted the label "neutral." Ideally, a subject who applies the label "neutral" to the second of two samples will interpret this as providing <u>zero</u> support for either hypothesis and hence will make no adjustment of the initial response. But if subjects do not recognize that "neutral" means "zero support," they may interpret the sample as evidence for another hypothesis, namely, that the machine is neither clearly normal nor clearly broken, and adjust toward the scale position (i.e., the midpoint) that best seems to signify

this third hypothesis.

In Experiment 2 the extent of debiasing was even more remarkable, particularly when one recalls the many previous unsuccessful efforts that have been directed at reducing conservatism in the Bayesian task. In evaluating this result, it is important to understand that there was nothing in the instructions that would have prevented subjects from continuing to weight information equally regardless of diagnosticity. That is, the subjects were only instructed to begin their judgment using the "stronger" sample; they were not instructed to give it more weight in the final judgment. Nevertheless, the net effect of the manipulation was for judgments to closely approximate optimal values. Whether this occurred because subjects intended to give the more important stimulus more weight is, of course, not clear. One might argue that the improved weighting pattern occurred due to unintentional primacy effects that were outside of the subjects' comprehension of the task. But it is worth noting that in natural judgment situations, people often "put first things first," considering those items that are deemed to be important before they consider other, less important items. Thus, one strategy for differential weighting may be exactly to attend to items in order of importance and to make smaller and smaller adjustments for items that are of lesser and lesser importance.

What Do Judges Do?

For more than 20 years, evidence has been accumulating that human judgments often seem to follow algebraic rules (cf. Anderson, 1974). The "averaging rule" for inference judgments is merely one case in point. But algebraic models of judgment have had limited appeal for some judgment researchers because they have only "as if" status: The data look as if they have been produced by application of an algebraic rule, but there is no theoretical necessity that the psychological processes of the judge in any way resemble "paper and pencil" algebraic manipulation.

The "debiasing" research reported here was based on a procedural theory of how people generate data that have algebraic patterns. The approach was based on the assumption that if a person's judgments in inference tasks look more like averages than like inferences, then at some point during the judgment process the person must be performing one or more operations that are nearer to those required for averaging than they are to those required for inferencing. Debiasing, then, must involve discovering those inappropriate judgment operations

and replacing them by operations that are better suited to inference.

Figure 7 outlines the major steps that are hypothesized to occur during judgment. In the first step, scanning, the judge merely assesses what information has been presented for judgment. Obviously, the details of the scanning step will depend on the task itself. In tasks such as that used in Experiment 1 where stimulus information is presented sequentially, scanning will be rudimentary since there is only one stimulus to scan. In simultaneous tasks, however, scanning will be more clearly distinguishable from other judgment operations. In some tasks (such as that used in Experiment 2) where the number of stimulus items is small and where there is no a priori reason to suppose that any particular item will be more important than any other item, scanning will include all available items, with order of scanning determined by stimulus formatting factors. In other cases, however, such as judging applications for graduate admission, some items may be consistently scanned before other items (i.e., GPA, GRE scores, etc.), and some items may be not scanned at all, at least not on the initial pass over an application (i.e., applicant's hobbies, past employment, etc.).

It is assumed that the scanning operation is primarily aimed at orienting the judge to the available information. Although the judge may develop a rough impression of the stimulus from scanning it, this impression will not in general be the final response. There may, of course, be exceptions to this rule. For example, if in scanning a graduate application, the judge notices that the candidate is clearly below standard on some critical factor, that application may be immediately rejected. Nevertheless, many experimental tasks implicitly or explicitly rule out such snap judgments by cautioning the judge against making overly hasty "end-responses."

Once the judge has scanned the available information, he or she is hypothesized to select an item to use as an "anchor point" (cf. Einhorn & Hogarth, 1982; Lopes & Johnson, 1982; Lopes & Oden, 1981; Tversky & Kahneman, 1974). If only one item has been presented, of course, that item must be the anchor. But if more than one item is available the "anchor stimulus" will generally be chosen because it seems relatively more important than the others. Such importance may reflect the a priori importance of the category to which the item belongs as, for example, GPA for graduate admissions. But it may also reflect diagnosticity within category as, for example, when items are selected by virtue of their being very extreme. Only if the various items seem equally important will the subject resort to ad hoc choice schemes such as, for example,

taking the items in order as they appear in the stimulus array.

Once an anchor has been chosen, it must be evaluated relative to the scale of judgment. This "valuation" operation may in some cases yield a quantity that serves directly as the initial judgment. For example, previous research on the inference task used in the present experiments suggests that subjects may simply anchor their judgment at a scale position that is proportional to the number of rejects in the first sample (Lopes, 1981). In other cases, however, the initial judgment may be somewhat less extreme than the scale value associated with the anchor stimulus. In these situations subjects act as though their initial judgment is a compromise between the value of the stimulus information and some internal, neutral "initial impression" (Anderson, 1967).

Once anchoring has been accomplished the subject must decide whether there are still important items left to be judged. If so, the process essentially reiterates, with the subject choosing which of the remaining items to consider next. As can be seen in the figure, the considerations at this point are exactly what they were at the time of choosing the anchor: If one of the remaining items is clearly more important than the others, the subject chooses it, otherwise an item is chosen arbitrarily, and the scale value of the chosen item is then determined.

The next step in the process is "adjustment" of the initial value in light of the new information. It is this step that is seen as being most crucial in determining the algebraic form of the overall judgment. In the case of "averaging" rules, the adjustment operation is assumed to involve two stages: (1) location of the new information on the scale of judgment relative to the initial judgment, and (2) adjustment of the initial judgment toward the new information. This produces a new judgment that lies between the first two values and is, in that sense, an average of the two. Other algebraic rules can also result from the adjustment stage, but the particulars of their respective adjustment processes differ in important ways. For example, multiplying can be seen as a form of serial fractionation (Lopes, 1976; Lopes & Ekberg, 1980) in which adjustments to the initial value are always downward (toward a zero-point on the response scale) and directly in proportion to the subjective value of the stimulus being adjusted for. In the same way, ratio responses, such as those produced by trained subjects in Experiment 2, can be seen as involving adjustments that reflect the degree to which new information supports or disconfirms the <u>qualitative</u> impact of previous information.

After each adjustment step, the judge is assumed to consider whether there are still important items left unaccounted for. If there are, the process is repeated with each new item of information leading to adjustment of the previous judgment, and with the adjustments ordinarily becoming smaller as the perceived importance of the new information becomes smaller in relation to previously considered information. At some point, however, either when the stimulus information runs out or when the subject judges that nothing important remains to be considered, the final response is made on whatever scale has been provided.

Changing What Judges Do

In order to debias human judgments, it is necessary to change the judgment process. But how? As Fischhoff (1982) pointed out, that depends on one's theory of why the judgments are biased. The earliest attempts at debiasing Bayesian inference were based generally on global notions of why the bias occurred: i.e., subjects were poorly motivated, or misunderstood the instructions, or refused to use the response scale properly. These causal models had in common that they implicitly assumed that the bias could be "fixed" without the necessity of knowing how the subject actually generated the biased judgment.

The present approach differs from these early methods in that it rests on an analysis of what the subject does when he or she erroneously produces an average rather than an inference. In this view, the judgment process is seen as comprising a set of procedures for scanning, selecting, analyzing and, finally, integrating stimulus information. The procedures are quantitative, but not numerical; computational but not arithmetical. The procedures function in such a way that judgments can be described fairly as the result of an averaging process, but the "algebra" is implicit in the subject's actions rather than explicit in conscious awareness.

In debiasing the present task, the first step was to understand exactly what subjects did in producing their biased judgments. Then it remained only to identify the faulty procedures and to replace these by similar--but normatively more appropriate--procedures. What is surprising is that for once the debiasing was even easier to do than to imagine, due in part to the fact that the training not only replaced "bad" procedures but, apparently, helped subjects to better understand the task.

Only one other debiasing method, that of Eils, Seaver, and Edwards (1977), has been as successful as the present method in improving the performance of

naive subjects in the Bayesian task. On the face of it, there are profound differences between these two successful approaches, not the least of which is that Eils et al. engineer the task to fit the subjects whereas the present approach engineers the subjects (or at least their judgment procedures) to fit the task. In a deeper sense, however, the two approaches have much in common because they are both based on an understanding of procedures that subjects use when they generate biased responses.

Eils et al. base their approach on the empirical observation that subjects, by whatever means, produce data that are more like averages than like inferences. They cleverly turn this "error" to their advantage by recasting the task so that subjects are asked to do what they do naturally and well, namely, averaging. It then requires only a simple mechanical transformation to convert the subjects' "average likelihood ratios" into "cumulative likelihood ratios."

The present approach is also based on an understanding of the averaging process, but the focus is shifted from the algebraic form of the data to the microstructure of the process that generates the data. The tacit assumption is that although subjects have access to components of the judgment process (and hence that they can control the sequence in which components are executed and the content on which they operate), they do not have assess to the algebraic implications of what their procedures do. Thus, there is little point in enjoining subjects to be less conservative, or to report their "true" probabilities, or to multiply rather than average.

Serious engineering in any domain rests on knowledge of the medium to be engineered. In the case of judgmental engineering, one must understand the judge as well as the judgment. People do some things well, other things badly; they have access to some processes and not to others. Building better judges requires that we know what people do and how they do it. After that, debiasing is easy.

References

Anderson, N. H. Information integration theory: A brief survey. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology, Volume 2. San Francisco: W. H. Freeman, 1974.

Anderson, N. H. Averaging model analysis of set-size effect in impression formation. Journal of Experimental Psychology, 1967, 75, 158-165.

Beach, L. R., Wise, J. A., & Barclay, S. Sample proportions and subjective probability revisions. Organizational Behavior and Human Performance, 1970, 5, 183-190.

Edwards, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), Formal representations of human judgment. New York: Wiley, 1968.

Eils, L. C., Seaver, D. A., & Edwards, W. Developing the technology of probabilistic inference: Aggregating by averaging reduces conservatism. Research Report 77-3, Social Science Research Institute, University of Southern California, August, 1977.

Einhorn, H. J., & Hogarth, R. M. A theory of diagnostic inference: I. Imagination and the psychophysics of evidence. Technical report, University of Chicago Graduate School of Business, June 1982.

Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases, Cambridge: Cambridge University Press, 1982.

Lopes, L. L. Averaging rules and adjustment processes: The role of averaging in inference. Technical report, Wisconsin Information Processing Program (WHIPP 13), Madison, WI, December 1981.

Lopes, L. L. Model-based decision and inference in stud poker. Journal of Experimental Psychology: General, 1976, 105, 217-239.

Lopes, L. L., & Ekberg, P. H. S. Test of an ordering hypothesis in risky decision making. Acta Psychologica, 1980, 45, 161-168.

Lopes, L. L., & Johnson, M. D. Judging similarity among strings described by hierarchical trees. Acta Psychologica, 1982, 51, 13-26.

Lopes, L. L., & Oden, G. C. Comparison of two models of similarity judgment. Acta Psychologica, 1980, 46, 205-234.

Marks, D. F., & Clarkson, J. K. An explanation of conservatism in the bookbag-and-pokerchips situation. Acta Psychologica, 1972, 36, 145-160.

Marks, D. F., & Clarkson, J. K. Conservatism as non-Bayesian performance: A reply to De Swart. Acta Psychologica, 1973, 37, 55-63.

Phillips, L. D., & Edwards, W. Conservatism in a simple probability inference task. Journal of Experimental Psychology, 1966, 72, 346-357.

Shanteau, J. C. An additive model for sequential decision making. Journal of Experimental Psychology, 1970, 85, 181-191.

Shanteau, J. C. Averaging versus multiplying combination rules of inference judgment. Acta Psychologica, 1975, 39, 83-89.

Shanteau, J. C. Descriptive versus normative models of sequential inference judgment. Journal of Experimental Psychology, 1972, 93, 63-68.

Troutman, C. M., & Shanteau, J. C. Inferences based on nondiagnostic information. Organizational Behavior and Human Performance, 1977, 19, 43-55.

Tversky, A., & Kahneman, D. Judgment under uncertainty: Heuristics and biases. Science, 1974, 185, 1124-1131.

Footnotes

The writing of this paper and the research reported therein were supported by the Office of Naval Research (Contract N0014-81-K-0069).

I am grateful to Rose Reed and Sandy Schneider for care in running the studies and to Gregg Oden for helpful comments on the manuscript.

TABLE 1

NUMBER OF ADJUSTMENT ERRORS

EXPERIMENT 1

PAIRS	CONTROL	TRAINED	MAXIMUM POSSIBLE
Weak-strong	0.40	0.42	24
Strong-weak	13.40	3.11	24
Diagonal	3.24	1.08	16

Note. Errors were scored for weak-strong cells and strong-weak cells only if there was actual adjustment and it was in the wrong direction. Errors were scored for diagonal cells only if there was room for adjustment and if there was adjustment in the wrong direction or no adjustment.

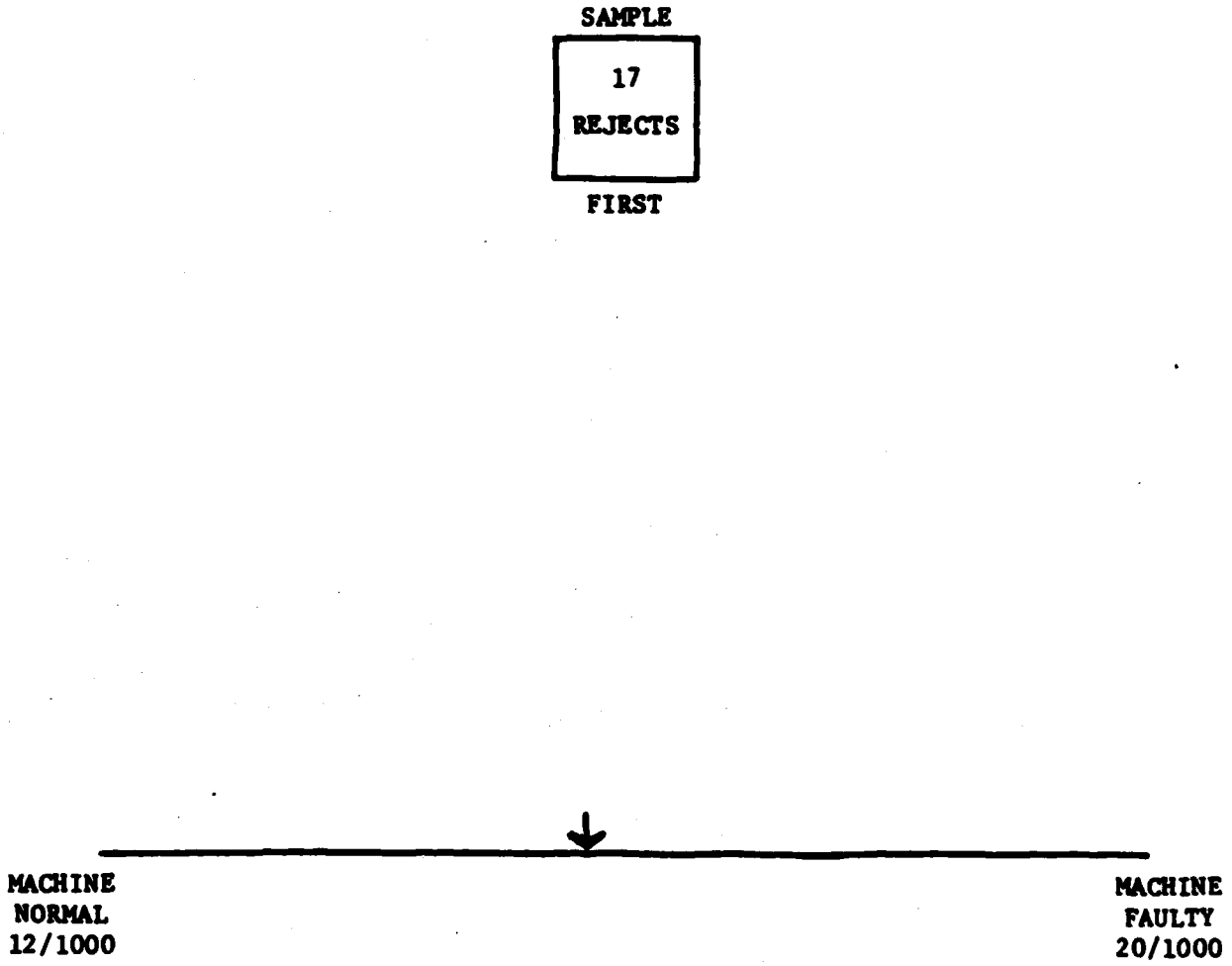

SAMPLE
17
REJECTS
FIRST
MACHINE
NORMAL
12/1000
MACHINE
FAULTY
20/1000

Figure 1

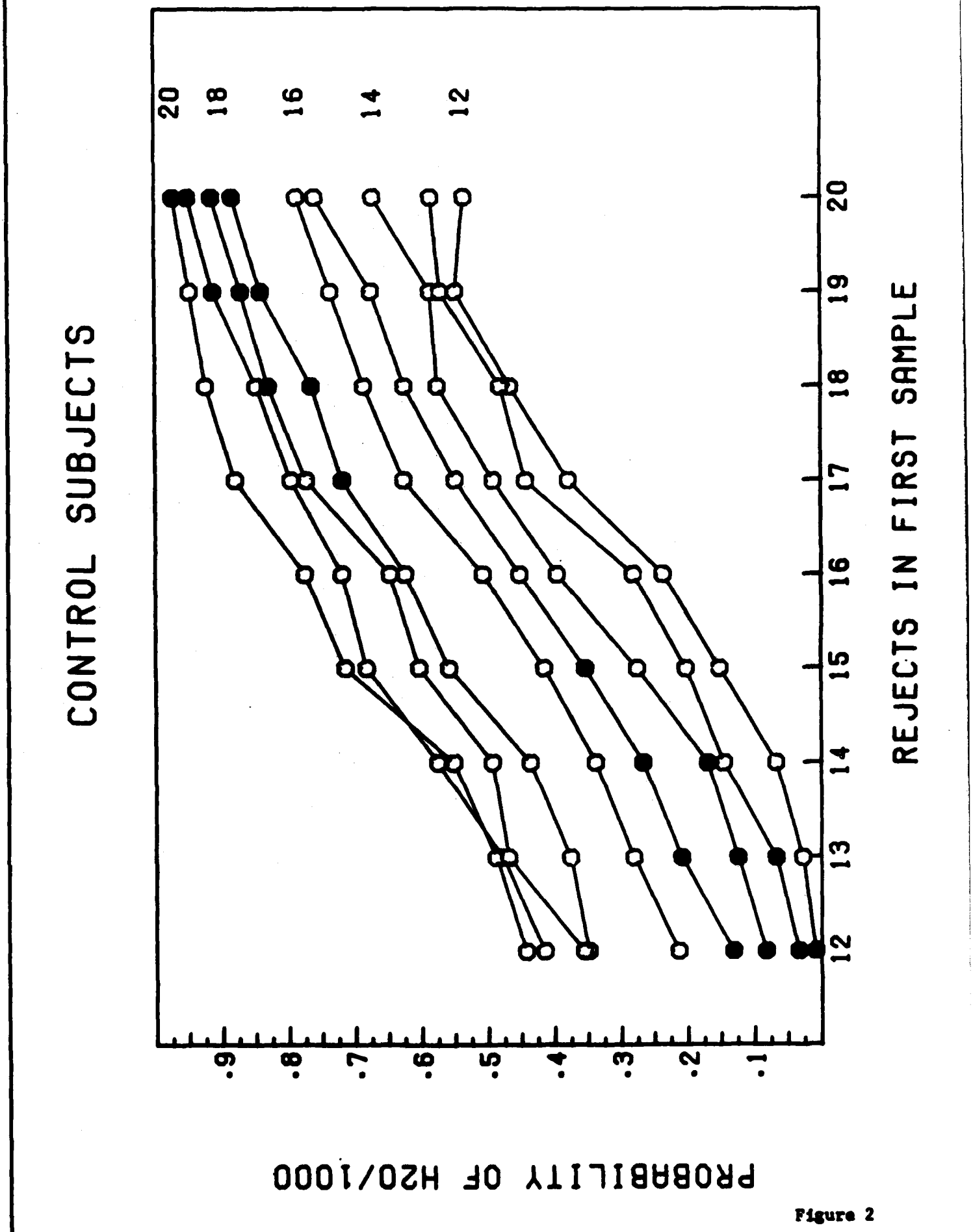

CONTROL SUBJECTS
20
18
16
14
12
PROBABILITY OF H20/1000
.9
.8
.7
.6
.5
.4
.3
.2
.1
12
13
14
15
16
17
18
19
20
REJECTS IN FIRST SAMPLE

Figure 2

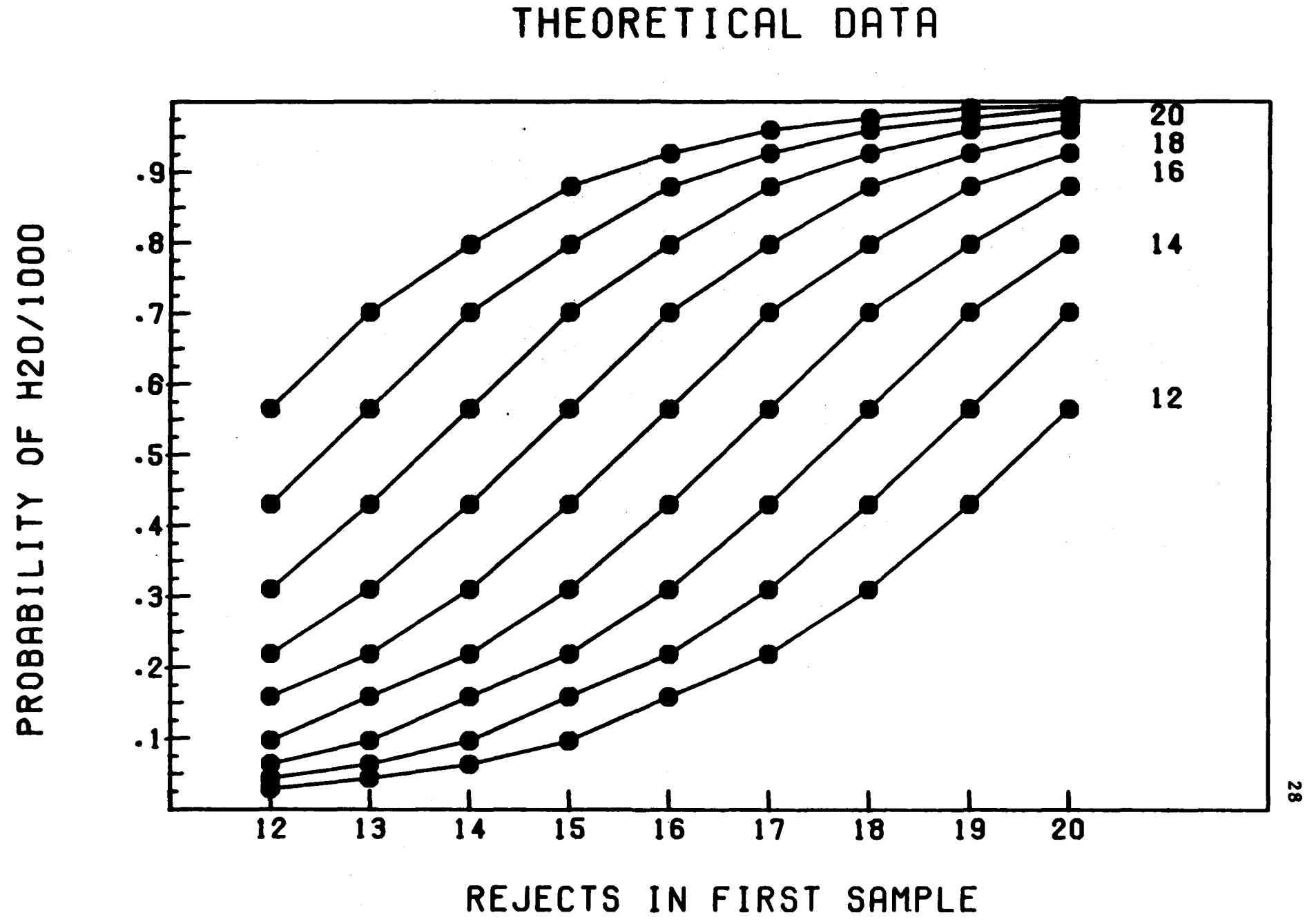

THEORETICAL DATA
PROBABILITY OF H20/1000
.9
.8
.7
.6
.5
.4
.3
.2
.1
20
18
16
14
12
12
13
14
15
16
17
18
19
20
REJECTS IN FIRST SAMPLE

Figure 3

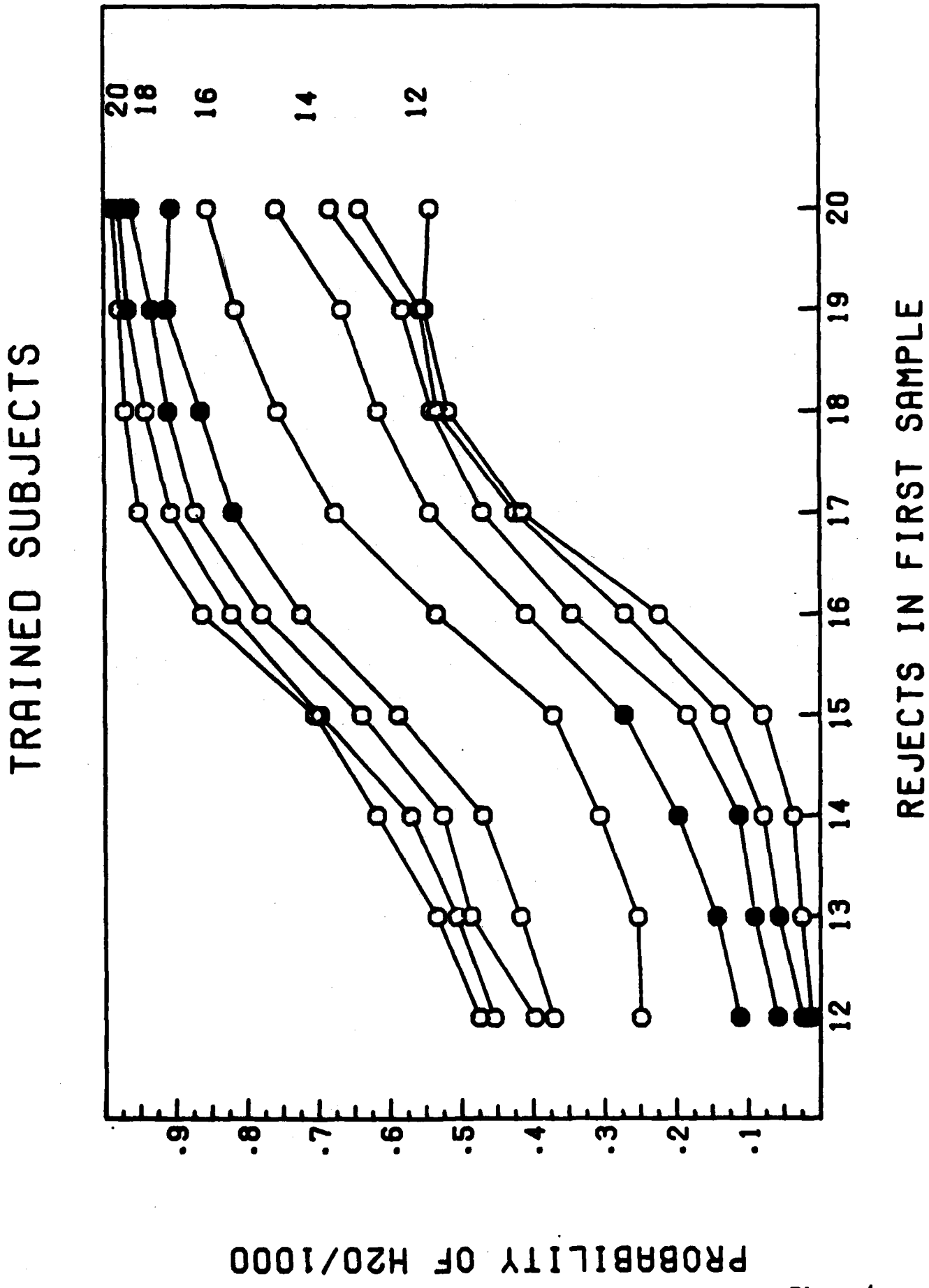

TRAINED SUBJECTS
REJECTS IN FIRST SAMPLE
PROBABILITY OF H20/1000
20
18
16
14
12
.9
.8
.7
.6
.5
.4
.3
.2
.1
20
19
18
17
16
15
14
13
12

Figure 4

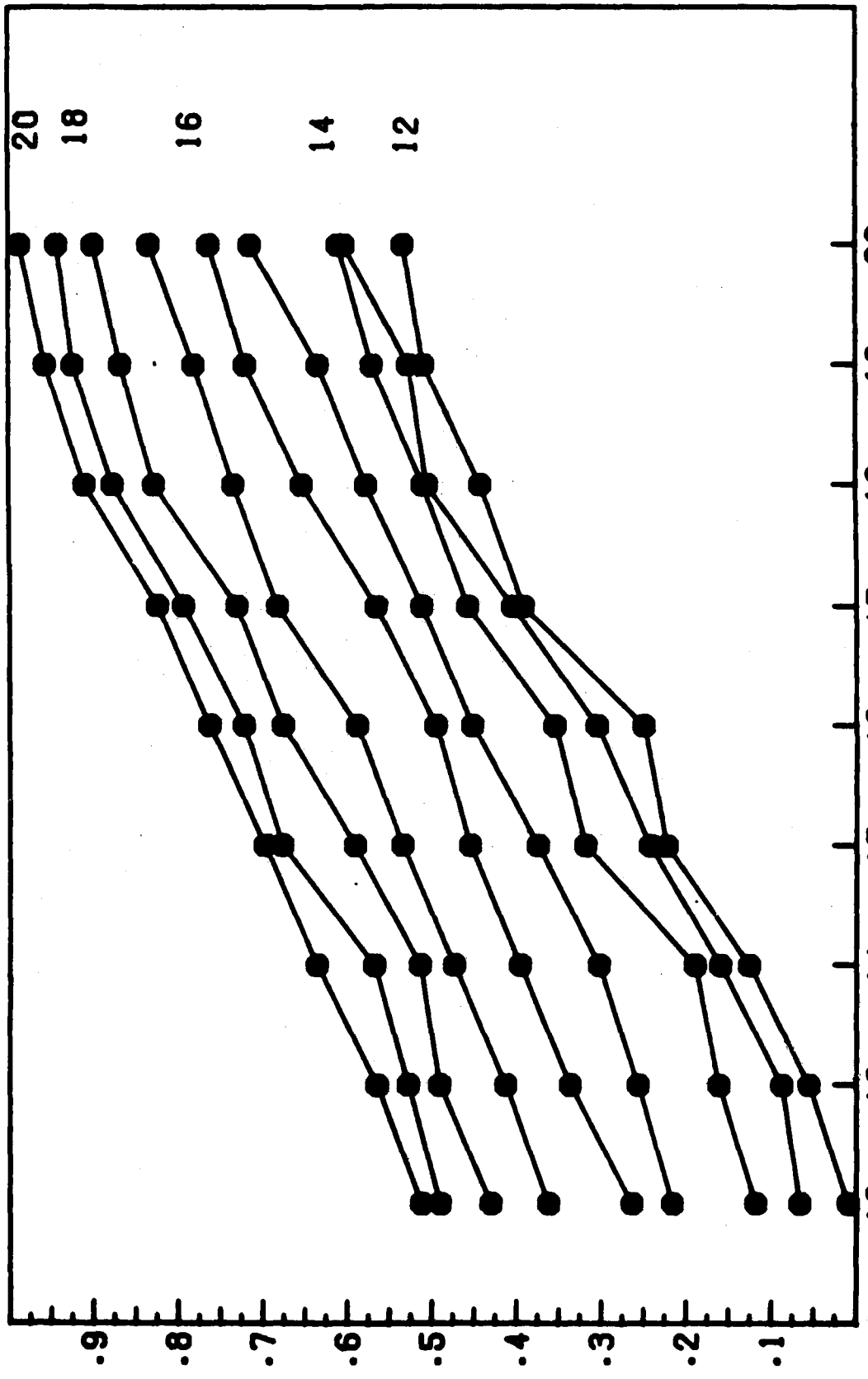

CONTROL SUBJECTS
20
18
16
14
12
.9
.8
.7
.6
.5
.4
.3
.2
.1
PROBABILITY OF H20/1000
12
13
14
15
16
17
18
19
20
REJECTS IN FIRST SAMPLE

Figure 5

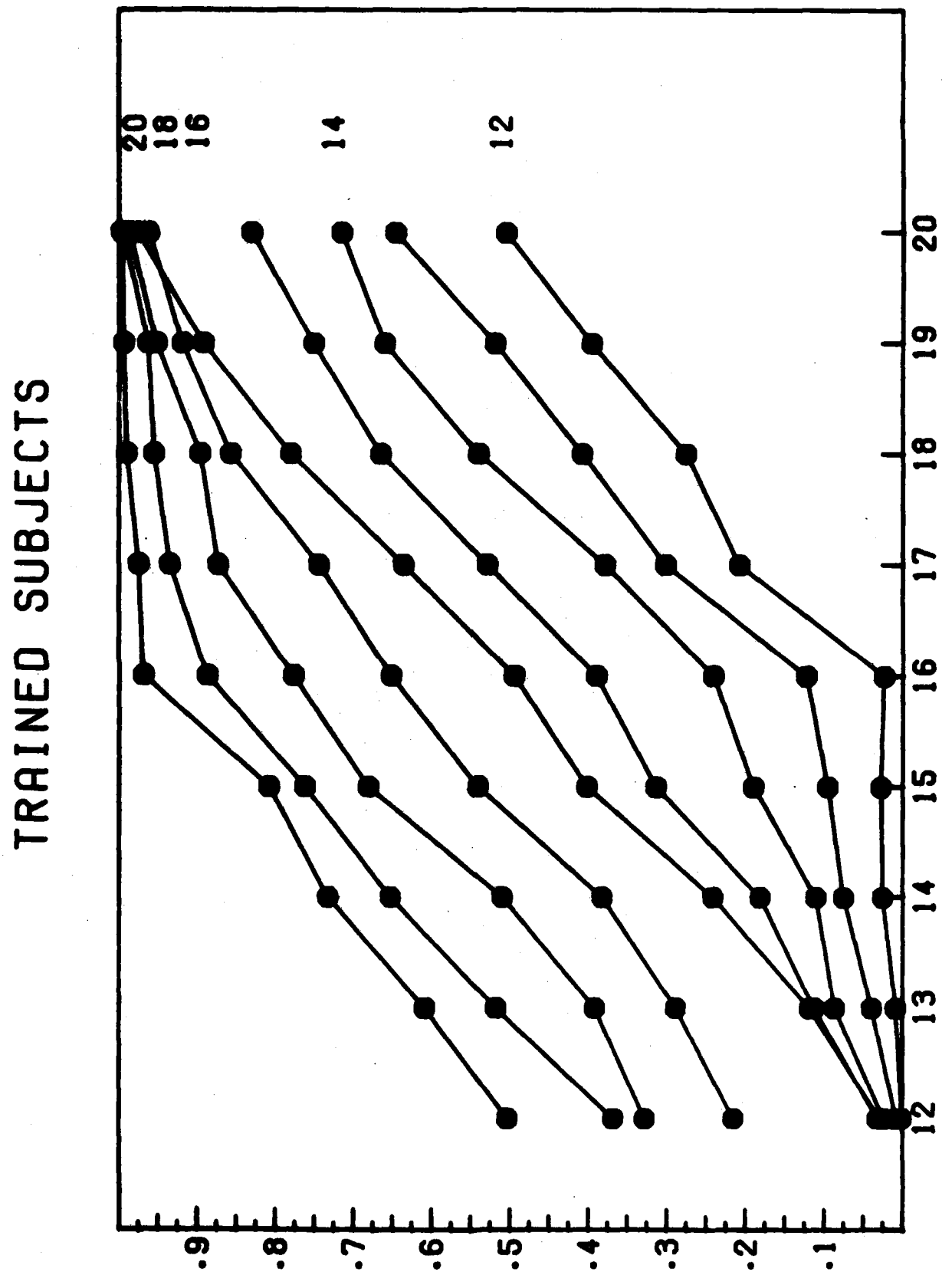

TRAINED SUBJECTS
PROBABILITY OF H20/1000
REJECTS IN FIRST SAMPLE
20
18
16
14
12
.9
.8
.7
.6
.5
.4
.3
.2
.1
12
13
14
15
16
17
18
19
20

Figure 6

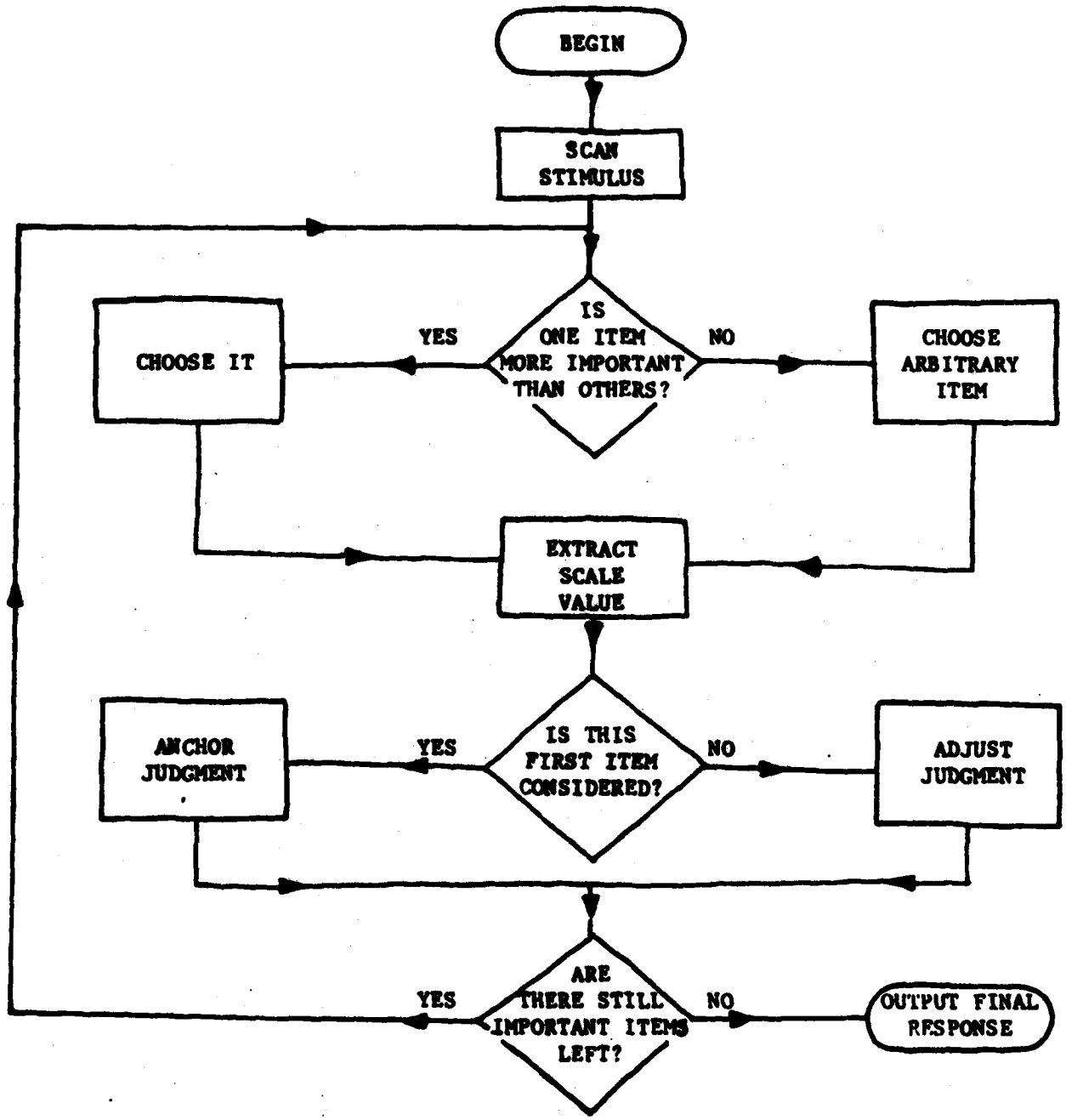

BEGIN
SCAN STIMULUS
IS ONE ITEM MORE IMPORTANT THAN OTHERS?
YES
NO
CHOOSE IT
CHOOSE ARBITRARY ITEM
EXTRACT SCALE VALUE
IS THIS FIRST ITEM CONSIDERED?
YES
NO
ANCHOR JUDGMENT
ADJUST JUDGMENT
ARE THERE STILL IMPORTANT ITEMS LEFT?
YES
NO
OUTPUT FINAL RESPONSE

Figure 7

DISTRIBUTION LIST

OSD

CAPT Paul R. Chatelier
Office of the Deputy Under Secretary
of Defense
OUSDRE (E&LS)
Pentagon, Room 3D129
Washington, D.C. 20301

Dr. Stuart H. Starr
Office of the Deputy Under Secretary
of Defense (C3I)
Pentagon
Washington, D.C. 20301

Department of the Navy

Engineering Psychology Programs
Code 442
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217 (5 cys)

Aviation & Aerospace Technology
Programs
Code 210
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Undersea Technology Programs
Code 220
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Electronics & Electromagnetics
Technology Programs
Code 250
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Communication & Computer Technology
Programs
Code 240
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Department of the Navy

Communication & Computer Technology
Programs
Code 240
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Tactical Development & Evaluation
Support Programs
Code 230
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Manpower, Personnel and Training
Programs
Code 270
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Mathematics Group
Code 411-MA
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Statistics and Probability Program
Code 411-S&P
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Information Sciences Division
Code 433
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

CDR K. Hull
Code 230B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Department of the Navy

Physiology & Neuro Biology Programs
Code 441B
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Special Assistant for Marine
Corps Matters
Code 100M
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Commanding Officer
ONREAST Office
ATTN: Dr. J. Lester
Barnes Building
495 Summer Street
Boston, MA 02210

Commanding Officer
ONRWEST Office
ATTN: Mr. R. Lawson
1030 East Green Street
Pasadena, CA 91106

Commanding Officer
ONRWEST Office
ATTN: Dr. E. Gloye
1030 East Green Street
Pasadena, CA 91106

Office of Naval Research
Scientific Liaison Group
American Embassy, Room A-407
APO San Francisco, CA 96503

Director
Naval Research Laboratory
Technical Information Division
Code 2627
Washington, D.C. 20375

Dr. Michael Melich
Communications Sciences Division
Code 7500
Naval Research Laboratory
Washington, D.C. 20375

Department of the Navy

Dr. Robert G. Smith
Office of the Chief of Naval
Operations, OP987H
Personnel Logistics Plans
Washington, D.C. 20350

CDR G. Worthington
Office of the Chief of Naval
Operations, OP-372G
Washington, D.C. 20350

Dr. W. Mehuron
Office of the Chief of Naval
Operations, OP 987
Washington, D.C. 20350

Dr. Andrew Technitzer
Office of the Chief of Naval
Operations, OP 952F
Naval Oceanography Division
Washington, D.C. 20350

Naval Training Equipment Center
ATTN: Technical Library
Orlando, FL 32813

Human Factors Department
Code N-71
Naval Training Equipment Center
Orlando, FL 32813

Dr. Alfred F. Smode
Training Analysis and Evaluation
Group
Naval Training Equipment Center
Code TAEG
Orlando, FL 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Dean of Research Administration
Naval Postgraduate School
Monterey, CA 93940

Department of the Navy

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, D.C. 20380

HQS, U.S. Marine Corps
ATTN: CCA40 (MAJOR Pennell)
Washington, D.C. 20380

Commanding Officer
MCTSSA
Marine Corps Base
Camp Pendleton, CA 92055

Chief, C^3 Division
Development Center
MCDEC
Quantico, VA 22134

Naval Material Command
NAVMAT 0722 - Rm. 508
800 North Quincy Street
Arlington, VA 22217

Commander
Naval Air Systems Command
Human Factors Programs
NAVAIR 340F
Washington, D.C. 20361

Commander
Naval Air Systems Command
Crew Station Design
NAVAIR 5313
Washington, D.C. 20361

Mr. Phillip Andrews
Naval Sea Systems Command
NAVSEA 0341
Washington, D.C. 20362

Commander
Naval Electronics Systems Command
Human Factors Engineering Branch
Code 81323
Washington, D.C. 20360

LCOL B. Hastings
Marine Corps Liaison Officer
Naval Coastal Systems Center
Panama City, FL 32407

Department of the Navy

Mr. Milon Essoglou
Naval Facilities Engineering Command
R&D Plans and Programs
Code 03T
Hoffman Building II
Alexandria, VA 22332

CDR Robert Biersner
Naval Medical R&D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Dr. Arthur Bachrach
Behavioral Sciences Department
Naval Medical Research Institute
Bethesda, MD 20014

CDR Thomas Berghage
Naval Health Research Center
San Diego, CA 92152

Dr. George Moeller
Human Factors Engineering Branch
Submarine Medical Research Lab
Naval Submarine Base
Groton, CT 06340

Head
Aerospace Psychology Department
Code L5
Naval Aerospace Medical Research Lab
Pensacola, FL 32508

Commanding Officer
Naval Health Research Center
San Diego, CA 92152

Dr. James McGrath
CINCLANT FLT HQS
Code 04E1
Norfolk, VA 23511

Navy Personnel Research and
Development Center
Planning & Appraisal Division
San Diego, CA 92152

Dr. Robert Blanchard
Navy Personnel Research and
Development Center
Command and Support Systems
San Diego, CA 92152

Department of the Navy

Mr. Stephen Merriman
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Dr. Julie Hopson
Human Factors Engineering Division
Naval Air Development Center
Warminster, PA 18974

Mr. Jeffrey Grossman
Human Factors Branch
Code 3152
Naval Weapons Center
China Lake, CA 93555

Human Factors Engineering Branch
Code 1226
Pacific Missile Test Center
Point Mugu, CA 93042

Dean of the Academic Departments
U.S. Naval Academy
Annapolis, MD 21402

Dr. S. Schiflett
Human Factors Section
Systems Engineering Test
Directorate
U.S. Naval Air Test Center
Patuxent River, MD 20670

Mr. John Quirk
Naval Coastal Systems Laboratory
Code 712
Panama City, FL 32401

CDR C. Hutchins
Code 55
Naval Postgraduate School
Monterey, CA 93940

Office of the Chief of Naval
Operations (OP-115)
Washington, D.C. 20350

Department of the Army

Mr. J. Barber
HQS, Department of the Army
DAPE-MBR
Washington, D.C. 20310

Department of the Army

Technical Director
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Director, Organizations and
Systems Research Laboratory
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333

Technical Director
U.S. Army Human Engineering Labs
Aberdeen Proving Ground, MD 21005

Department of the Air Force

U.S. Air Force Office of Scientific
Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, D.C. 20332

Chief, Systems Engineering Branch
Human Engineering Division
USAF AMRL/HES
Wright-Patterson AFB, OH 45433

Dr. Earl Alluisi
Chief Scientist
AFHRL/CCN
Brooks AFB, TX 78235

Foreign Addresses

Dr. Daniel Kahneman
University of British Columbia
Department of Psychology
Vancouver, BC V6T 1W5
Canada

Dr. Kenneth Gardner
Applied Psychology Unit
Admiralty Marine Technology
Establishment
Teddington, Middlesex TW11 OLN
England

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
Post Office Box 2000
Downsview, Ontario M3M 3B9
Canada

Foreign Addresses

Dr. A. D. Baddeley
Director, Applied Psychology Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF
England

Other Government Agencies

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314 (12 cys)

Dr. Craig Fields
Director, System Sciences Office
Defense Advanced Research Projects
Agency
1400 Wilson Blvd.
Arlington, VA 22209

Dr. Lloyd Hitchcock
Federal Aviation Administration
ACT 200
Atlantic City Airport, NJ 08405

Dr. M. Montemerlo
Human Factors & Simulation
Technology, RTE-6
NASA HQS
Washington, D.C. 20546

Dr. J. Miller
Florida Institute of Oceanography
University of South Florida
St. Petersburg, FL 33701

Other Organizations

Dr. Robert R. Mackie
Human Factors Research Division
Canyon Research Group
5775 Dawson Avenue
Goleta, CA 93017

Dr. Amos Tversky
Department of Psychology
Stanford University
Stanford, CA 94305

Other Organizations

Dr. H. McI. Parsons
Human Resources Research Office
300 N. Washington Street
Alexandria, VA 22314

Dr. Jesse Orlansky
Institute for Defense Analyses
1801 N. Beauregard Street
Alexandria, VA 22311

Professor Howard Raiffa
Graduate School of Business
Administration
Harvard University
Boston, MA 02163

Dr. T. B. Sheridan
Department of Mechnical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Arthur I. Siegel
Applied Psychological Services, Inc.
404 East Lancaster Street
Wayne, PA 19087

Dr. Paul Slovic
Decision Research
1201 Oak Street
Eugene, OR 97401

Dr. Harry Snyder
Department of Industrial Engineering
Virginia Polytechnic Institute and
State University
Blacksburg, VA 24061

Dr. Robert T. Hennessy
NAS - National Research Council (COHF)
2101 Constitution Ave., N.W.
Washington, D.C. 20418

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Other Organizations

Dr. Robert Williges
Dept. of Industrial Engineering & OR
Virginia Polytechnic Institute
and State University
130 Whittemore Hall
Blacksburg, VA 24061

Dr. Meredith P. Crawford
American Psychological Association
Office of Educational Affairs
1200 17th Street, N.W.
Washington, D.C. 20036

Dr. Deborah Boehm-Davis
General Electric Company
Information Systems Programs
1755 Jefferson Davis Highway
Arlington, VA 22202

Dr. Ward Edwards
Director, Social Science Research
Institute
University of Southern California
Los Angeles, CA 90007

Dr. Robert Fox
Department of Psychology
Vanderbilt University
Nashville, TN 37240

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
455 West Lindsey
Norman, OK 73069

Dr. Kenneth Hammond
Institute of Behavioral Science
University of Colorado
Room 201
Boulder, CO 80309

Dr. James H. Howard, Jr.
Department of Psychology
Catholic University
Washington, D.C. 20064

Dr. William Howell
Department of Psychology
Rice University
Houston, TX 77001

Other Organizations

Dr. Christopher Wickens
University of Illinois
Department of Psychology
Urbana, IL 61801

Mr. Edward M. Connelly
Performance Measurement
Associates, Inc.
410 Pine Street, S.E.
Suite 300
Vienna, VA 22180

Prof. Michael Athans
Room 35-406
Massachusetts Institute of Technology
Cambridge, MA 02139

Dr. Edward R. Jones
Chief, Human Factors Engineering
McDonnell-Douglas Astronautics
Company
St. Louis Division
Box 516
St. Louis, MO 63166

Dr. Babur M. Pulat
Department of Industrial Engineering
North Carolina A&T State University
Greensboro, NC 27411

Dr. A. K. Bejczy
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91125

Dr. Stanley N. Roscoe
New Mexico State University
Box 5095
Las Cruces, NM 88003

Mr. Joseph G. Wohl
Alphatech, Inc.
3 New England Industrial Park
Burlington, MA 01803

Dr. Rex Brown
Decision Science Consortium
Suite 721
7700 Leesburg Pike
Falls Church, VA 22043

Other Organizations

Dr. Wayne Zachary
Analytics, Inc.
2500 Maryland Road
Willow Grove, PA 19090

Dr. William R. Uttal
Institute for Social Research
University of Michigan
Ann Arbor, MI 48109

Dr. Richard Pew
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

Dr. Hillel Einhorn
University of Chicago
Graduate School of Buriness
1101 E. 58th Street
Chicago, IL 60637

Dr. David J. Getty
Bolt Beranek & Newman, Inc.
50 Moulton Street
Cambridge, MA 02238

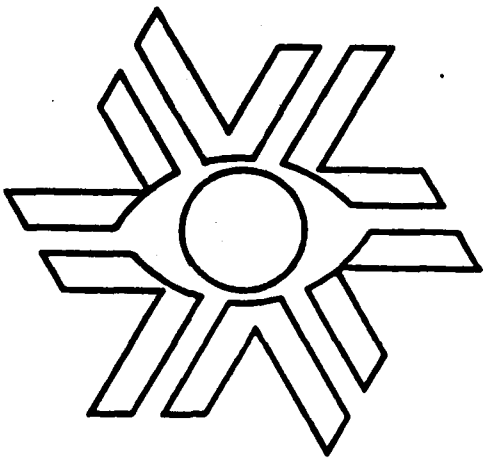
Dr. Douglas Towne
University of Southern California
Behavioral Technology Laboratory
3716 S. Hope Street
Los Angeles, CA 90007

Dr. John Payne
Duke University
Graduate School of Business
Administration
Durham, NC 27706

Dr. Baruch Fischhoff
Decision Research
1201 Oak Street
Engene, OR 97401

Dr. Andrew P. Sage
University of Virginia
School of Engineering and Applied
Science
Charlottesville, VA 22901

1. Gregg C. Oden & Dominic W. Massaro - Integration of place and voicing information in identifying synthetic stop-consonant syllables. July 1977.
2. Lola L. Lopes & Gregg C. Oden - Judging similarity among kinship terms. October 1977.
3. Gregg C. Oden - On the use of semantic constraints in guiding syntactic analysis. January 1978.
4. Howard J. Kallman & Dominic W. Massaro - Similarity effects in backward recognition masking. May 1978.
5. Steven J. Lupker & Dominic W. Massaro - Selective perception without confounding contributions of decision and memory. May 1978.
6. Gregg C. Oden & James L. Spira - Influence of context on the activation and selection of ambiguous word senses. August 1978.
7. Lola L. Lopes & Per-Hakan S. Ekberg - Serial fractionation in risky choice: Test of an analog process for multiplicative judgment. August 1978.
8. Marcia A. Derr & Dominic W. Massaro - The contribution of vowel duration, F_0 contour, and frication duration as cues to the /juz/ - /jux/ distinction. September 1978.
9. Dominic W. Massaro & Gregg C. Oden - Evaluation and integration of acoustic features in speech perception. September 1978.
10. Gregg C. Oden & Lola Lopes - Kin search: Answering questions about relations among relatives. September 1979.
11. Gregg C. Oden & Lola L. Lopes - On the internal structure of fuzzy subjective categories. September 1980.
12. Lola L. Lopes - Decision Making in the Short Run. October 1980.
13. Lola L. Lopes - Averaging Rules and Adjustment Processes: The Role of Averaging in Inference. December 1981.
14. Gregg C. Oden - Integration of Linguistic Information in Language Comprehension. April 1982.
15. Lola L. Lopes - Procedural Debiasing. October 1982.



University of Wisconsin-Madison